



Atomic Energy
Control Board

P.O. Box 1046
Ottawa, Canada
K1P 5S9

Commission de contrôle
de l'énergie atomique

C.P. 1046
Ottawa, Canada
K1P 5S9

INFO-0266

"ALIVE" SEARCHES AS COMPLEMENTING
DEATH SEARCHES IN THE EPIDEMIOLOGICAL
FOLLOW-UP OF ONTARIO MINERS

by

M.E. Fair, H.B. Newcombe,
P. Lalonde and C. Poliquin
Statistics Canada

A research report prepared for the
Atomic Energy Control Board
Ottawa, Canada

February 1988

88-06136/

"ALIVE" SEARCHES AS COMPLEMENTING DEATH SEARCHES
IN THE EPIDEMIOLOGICAL FOLLOW-UP OF ONTARIO MINERS

A report prepared by Martha E. Fair, Howard B. Newcombe, Pierre Lalonde and Christine Poliquin, Occupational and Environmental Health Unit, Vital Statistics and Disease Registries Section, Health Division, Statistics Canada, under contract to the Atomic Energy Control Board.

ABSTRACT

"Alive" searches have been used to complement the "death" searches, in a study of the mortality experience of a cohort of Ontario miners. The purpose has been to develop a way of distinguishing between those cohort members who are "confirmed" alive at a given time, and those who are "lost to follow-up".

A total of 30,000 Workers' Compensation Board (WCB) records with valid Social Insurance Numbers (SIN) were used to search the income tax files by computer over two consecutive years (1977 and 1978) representing nearly 27 million tax returns. These tax file searches using SIN numbers have provided information on the procedures to be used in, and likely success of, the corresponding searches of the tax files that could be carried out where the SIN number is not available on the work records. The latter kind of search would have to be based on names, birth dates and such, and would be probabilistic in nature.

The results of the study were as follows. After the initial death search, it was found that 7.5% of the cohort had died in Canada and their records were found on the Mortality Data Base. The remaining 92.5% has been "assumed alive" in the earlier analysis. After the "alive" follow-up using the income tax file, one was able to confirm that 89.1% of the "assumed alive" had filed an income tax after the study period. Thus only 3.4% of the cohort remained untraced. Among those there could be an admixture of deaths outside of Canada, persons who have moved outside of Canada and are still alive, and/or persons alive within Canada, but who have not filed an income tax return.

Probabilistic tax linkages were also carried out, based on names, birth dates, marital status and such. The discriminating power of these identifiers was shown to be high provided the date of birth was present on the work record, and especially high wherever one or both full given names were used instead of just the initials. Probabilistic linkages with the tax files, in the absence of the SIN number, should therefore be feasible on a routine basis, and reasonably accurate.

This study has indicated that the procedures developed are useful for purposes of "alive" follow-up, for evaluation of the quality of the Mortality Data Base-cohort death file searches, for improving the accuracy of analytical results of epidemiological studies, and for reducing the cost and labour of resolving doubtful death searches.

RÉSUMÉ

Les recherches concernant les mineurs vivants ont servi de complément aux recherches concernant les mineurs décédés, dans le cadre d'une étude de mortalité menée auprès d'une cohorte de mineurs de l'Ontario, afin de trouver une façon d'établir une distinction entre les membres de la cohorte qui sont "confirmés" vivants à un moment donné et ceux qui sont "perdus" au cours du suivi.

En tout, 30 000 enregistrements de la Commission des accidents de travail comportant un numéro d'assurance sociale valide ont servi à effectuer des recherches informatisées dans les fichiers d'impôt sur le revenu de deux années consécutives (1977 et 1978), ce qui représente près de 27 millions de déclarations d'impôt. Ces recherches réalisées à partir du numéro d'assurance sociale ont fourni des données sur les procédures à utiliser dans les recherches correspondantes des fichiers d'impôt lorsque le numéro d'assurance sociale ne figure pas sur le dossier de travail des employés, d'une part, et sur leurs chances de réussite, d'autre part. Ces recherches s'effectueraient à partir du nom, de la date de naissance, etc., et seraient de nature probabiliste.

Suivant les résultats de l'étude, les recherches initiales concernant les mineurs décédés ont révélé que 7,5 pour 100 de la cohorte étaient morts au Canada et que leur enregistrement se trouvait dans la Base de données sur la mortalité. Quant aux autres membres de la cohorte (92,5 pour 100), on a d'abord supposé qu'ils vivaient toujours. Après le suivi auprès des mineurs «supposés vivants», réalisé à l'aide des fichiers d'impôt sur le revenu, on a pu confirmer que 89,1 pour 100 des mineurs «supposés vivants» avaient rempli une déclaration d'impôt après la période d'étude. Ainsi, seuls 3,4 pour 100 des membres de la cohorte demeurent introuvables pour différentes causes : ils pourraient être décédés hors du Canada, avoir déménagé à l'étranger et être toujours vivants, ou habiter toujours au pays, mais ne pas avoir rempli de déclaration d'impôt.

On a également procédé à des couplages probabilistes avec les fichiers d'impôt à partir du nom, de la date de naissance, de l'état civil, etc. Ces identificateurs ont permis de retrouver bon nombre de mineurs si leur date de naissance figurait sur leur dossier de travail, et davantage si l'un des deux prénoms ou les deux au long apparaissaient au lieu des simples initiales. Par conséquent, on devrait pouvoir réaliser régulièrement et de façon assez exacte des couplages probabilistes avec les fichiers d'impôt, si le numéro d'assurance sociale n'apparaît pas sur le dossier de travail des employés.

L'étude a montré que les procédures établies sont utiles pour assurer le suivi auprès des mineurs vivants, à évaluer la qualité du couplage des dossiers de travail des employés avec la Base de données sur la mortalité, à améliorer l'exactitude des résultats analytiques d'études épidémiologiques et à économiser les ressources financières et humaines pour ce qui est de mener à terme des recherches incertaines concernant les mineurs décédés.

DISCLAIMER

The Atomic Energy Control Board is not responsible for the accuracy of the statements made or opinions expressed in this publication and neither the Board nor the author assumes liability with respect to any damage or loss incurred as a result of the use made of the information contained in this publication. The opinions expressed in this paper are those of the authors and do not necessarily represent the views of Statistics Canada.

PREFACE

The present version of the report on the "alive" follow-up has been prepared following production of final data from the relevant linkages pertaining to the Ontario Miners study.

The procedures described here are those actually followed, and are not always the same as those discussed tentatively in an earlier planning document (the "Interim Report") which was written prior to initiation of much of the work. A number of potential data sources (Canada Pension Plan, Old Age Security and Guaranteed Income supplement, Ontario Drug Benefit Eligibility file, Ontario Driver's license, census numerations) were examined before the present feasibility study was initiated using the income tax data files. Readers wanting more detail of the earlier tentative plans are referred to:

"Ontario Miners "Alive" follow-up Feasibility Study -- the 1983/84 Component".

Occupational and Environmental Health Research Unit, Vital Statistics and Health Status Section, Health Division, Statistics Canada, Ottawa, March 1984.

Those who seek information about the vital status of a cohort of people are apt to vary widely in the "quality" of the product they require. For epidemiologists, it is frequently important that doubtful death linkages be resolved, and that any circumstances which might preclude a death linkage, such as migrations out of the country, be identified wherever possible. For other statistical purposes, a crude estimate of the proportion who have died may sometimes be sufficient. The distinction between these two sorts of users is important, because the associated cost and effort will be different.

Epidemiologists doing cohort studies, particularly where the setting of regulatory standards and development of compensation criteria are involved, often fall into the former category. Often large numbers of individuals are involved in study groups looking at the long-term health effects of low doses of radiation. High follow-up is desired so that no assumptions need be made in computing mortality rates among the cohort being studied.

The present study will be of particular interest to those epidemiologists who find it desirable to determine the vital status at the end of the potential period of follow-up for all persons and where intensive individual follow-up is to be conducted and/or manual review of death searches is to be carried out. More specifically, it will be of particular interest:

- (i) to reduce the number and cost of manually resolving "doubtful" death searches;
- (ii) to reduce the cost and labour of intensive active individual follow-up of a complete cohort or a sample of the cohort;
- (iii) to develop the methodology and to use existing Canadian data sources for a wide variety of long-term follow-up studies.

TABLE OF CONTENTS	PAGE
ABSTRACT	i
PREFACE	iii
A. Introduction	1
B. Available files	2
C. Plan of the study	2
D. Procedures for the numeric linkages	3
E. Results from the numeric linkages	3
F. Procedures for the probabilistic linkages	4
G. Results from the probabilistic linkages	6
H. Discussion of the value of "alive" follow-up	6
TEXT TABLES	
1. Numeric of miners WCB records used for the "alive" follow-up	8
2. Vital Status of the 30,000 miners as indicated by the death searches	9
3. Characteristics of the taxation files used to verify the vital status of the 30,000 miners	10
4. Results from matching with the 1977 taxation file alone	11
5. Results from matching with both the 1977 and 1978 taxation files, after manually resolving the conflict between tax years	12
6. Final disposition of the WCB-tax conflicts and non-confirmations shown in the previous Table	13
7. Vital status as judged prior to and following the taxation searches and manual resolutions summary	14
8. Frequencies of good matches, false matches and non-matches from the probabilistic "alive" searches	15
9. Accuracy of the probabilistic "alive" searches with different amounts of identifying information	16

	PAGE
APPENDICES AND APPENDIX TABLES	
A. Vital status as judged prior to and following the taxation searches and manual resolutions -- detail	A
Appendix Table A-1 Vital status as judge prior to and following the taxation searches and manual resolutions -- detail	A3
B. Procedures for the probabilistic linkages -- detail	B
Appendix Table B-1 Identifiers compared between the Workers Compensation Board records and those of the Taxation Files	B3
Appendix Table B-2 Weights assigned to the comparison outcomes -- first comparisons	B4
Appendix Table B-3 Weights for comparisons with PROVINCE OF RESIDENCE as given on the 1977 Taxation File -- calculated after the linkage versus actually used	B8
Appendix Table B-4 Weights for comparisons with LOCALITY OF RESIDENCE as given on the 1977 Taxation File -- calculated after the linkage versus actually used	B9
Appendix Table B-5 Weighting for comparison of birth years	B11
Appendix Table B-6 Weights in need of further study -- SURNAME and SURNAME NYSIIS	B12
Appendix Table B-7 Weights in need of further study -- FIRST GIVEN NAMES	B13
C. Accuracy of the probabilistic linkages -- detail	C
Appendix Table C-1 Accuracy of the probabilistic alive searches with different amounts of identifying information -- detail	C2
D. Some essential features of mortality follow-up studies among industrial populations	D
E. Dealing with persons "Lost to follow-up"	E
F. Statistics Canada Policy on record linkage	F

A. INTRODUCTION

An ambiguity exists wherever the mortality experience of a study cohort is judged solely from the results of a death search. A proportion of those members not known to be dead may in reality be "lost to follow-up", either because they have migrated out of the country or perhaps because of name changes. Moreover, where the personal identifying information is poor on either a search record or a matching death record, there will be uncertainty concerning the possible deceased status of the cohort member. To resolve the ambiguity that arises from these various sources, some kind of "alive" search is needed to serve as a counterpart to the "death" search. Centralized death searches for the purpose of health studies are now routine in this country; it remains to be demonstrated that the corresponding "alive" searches are also feasible.

Resources that would facilitate both types of search are well developed in Canada. The centralized Mortality Data Base (MDB) exists in machine readable form and is consolidated over all provinces and territories and over spans of years. In addition, files that could verify the alive status of an individual are numerous at all levels of government and elsewhere (e.g. municipal censuses, tax rolls and utility billings, voters lists at all three levels of government, telephone directories, eligibility lists for free drugs, Old Age Security and Canada Pension Plan files, drivers licences, passport applications, federal income tax files, etc.) Many of these sources are now computerized.

For the present study it was decided to do an "alive" follow-up of a cohort of Ontario miners who had already been the subjects of automated death searches. Income tax files were chosen to permit the complementary "alive" search. Because many of the Workers' Compensation Board (WCB) records for the miners carried the Social Insurance Number (SIN) the initial searching has been for matches using this number. The experience thus gained is of value for the design of procedures that could be used in similar searches that do not require the SIN and would be based instead on probabilistic linkages using names, birth dates and such.

B. AVAILABLE FILES

Approximately 50,000 Workers' Compensation Board (WCB) records pertaining to Ontario miners were available for the study. Also available were the Canadian income tax records for the years 1977 and 1978 numbering close to 27,000,000 for the two years together.

These files are capable of being linked with each other in either of two ways, a) by searching for exact matches of the SIN numbers, or b) by comparing the names, birth dates and such, using probabilistic methods to distinguish the similarities that are unlikely to have occurred together by chance. Of the WCB records, approximately 30,000 contain the SIN number (Table 1). Those having acceptable SIN numbers include over 2,000 that have either been positively linked with a death record or at least are highly likely to have been correctly linked with one (Table 2). Of the taxation records, virtually all have acceptable SIN numbers (Table 3). Both files contain in common the names, birth date, sex, and province of residence of the individual, although some of these identifiers may be lacking from certain of the WCB records.

C. PLAN OF THE STUDY

Work on this study fell into two major parts or phases. In the first phase the WCB records that had SIN numbers were linked with any taxation records that had precisely matching SINs. In the second phase, the experience gained in the first phase was used to design the procedures for a probabilistic linkage between the WCB and the tax records, independent of the SIN and using instead the names, birth dates and such.

The numeric linkages using the SIN number were technically simple. These were carried out first with the 1977 tax file alone, and subsequently with the 1977 and 1978 tax files together after resolving any apparent conflicts between the two tax files with respect to the reported vital status of the tax payer.

The probabilistic procedures were tested using the WCB records that had already been positively linked to tax records by means of the SIN numbers. In this way, any failures of known potentially linkable pairs of records to link correctly and strongly by the probabilistic methods could be readily detected. Moreover, the extent of the dependence of the probabilistic approach on the availability of the various identifiers could be assessed, thereby providing guidance when future decisions are to be made concerning how much identifying information is needed to ensure a given level of accuracy of linkage.

D. PROCEDURES FOR THE NUMERIC LINKAGES

The WCB initiated searches of the tax files were carried out in numeric sequence in order of the Social Insurance Number. Only exact matches on the SIN number were sought.

As a refinement, the present plan included two separate numeric linkages of the WCB records, first with a single year of tax records (1977), and then with two successive years (1977 and 1978) taken together. This feature was intended to result in improved evidence of the alive status of an individual, plus better coverage of the WCB cohort since some individuals could have missed sending in tax return for one of the years. It was viewed also as an aid to spotting errors in the reporting and coding of dead-or-alive status on the tax records.

When the pairs of records had been matched on the SIN number, the other identifiers were also compared for the purpose of estimating by probabilistic methods the likelihood of a correct match. This likelihood was almost always high; but where it was not, the pair were checked visually to determine the cause of the seeming conflict.

E. RESULTS FROM THE NUMERIC LINKAGES

For the majority of the 30,000 Ontario miners, the two numeric linkages with the tax files provided further evidence of the dead-or-alive status as judged earlier from the death searches alone. Those miners who were believed to have been dead by the end of 1977 rarely filed tax returns in the 1977-78 period. Conversely, those who were assumed to be alive rarely failed to do so. However it is the seeming conflicts between the two sorts of evidence that are of special interest, plus the instances in which there is no evidence either of death or of life. These included the miners who were believed to be dead but for whom tax returns had indeed been filed, plus those for whom there was neither a known death record nor any tax record.

The first of the two "alive" searches yielded 80 apparent conflicts, and indicated that there was no evidence of either sort for 1,465 of the miners (Table 4). The second "alive" search reduced the apparent conflicts to just 14 cases, all of which were resolved manually (Table 5). At this stage the previously doubtful death linkages could also be resolved, establishing a final figure of 2,254 deaths to the end of 1977, and confirming that a group of 1,010 were lost to follow-up by present methods.

The second of the two numeric linkages was preceeded by a manual resolution of the apparent conflicts between the tax files themselves, and was followed by the further visual resolution of the remaining conflicts between the WCB records and the tax records. There were four main sources of trouble. These had to do:

- (a) with the incorrect assignment of a "female" sex code on the tax records (4 cases),
- (b) with persons who were wrongly coded "dead" when they were alive (5 cases),
- (c) with those who did in fact die but were coded on their tax records as "alive" when they had died prior to the end of 1977 (9 cases), and
- (d) with those who were coded "dead" in both 1977 and 1978, where the death had actually occurred in 1978 but prior to submission of the 1977 tax return (38 cases).

The final dispositions are shown in Table 6, and the data are summarized in Table 7. Details are contained in Appendix A and its Table A.1.

Identification of the "lost to follow-up" category would not have been possible from the death searches alone. This group amounts to between 3 and 4 per cent of the individuals who had previously been "presumed alive", and collectively they approached 50 per cent of the size of the "known dead" group.

F. PROCEDURES FOR THE PROBABILISTIC LINKAGES

A further purpose of the numeric linkages was to determine whether comparable searches of the tax files could be carried out successfully in a sequence based on the surnames, when the SIN numbers are not available on the work records. These searches and linkages would necessarily be probabilistic in nature.

Such "nominal" linkages with the tax files were carried out using the same WCB records that had been employed earlier in the numeric searches. The work records selected were those that were already known to be correctly linkable with particular tax records. Thus there would be no difficulty in judging the validity of the probabilistic linkages based on the names and such. Moreover, the prior numeric linkages served to provide information needed for refining the nominal linkage procedure. In particular, it is desirable to know the frequencies with which each of the identifiers agree, disagree, or partially agree and partially disagree, among pairs of records that have been correctly matched. The reason is that, when calculating the probabilities, one must ensure that too much negative weight is not given to discrepancies which are relatively common, and that too little negative weight is not given to the rare discrepancies when they do occur.

The identifiers that were used for the probabilistic linkages were:

SURNAME, including a phonetically coded (NYSIIS) version of it,

INITIALS, i.e. the 1st and 2nd, which were first compared directly and then cross-compared wherever they both disagreed on the straight comparison,

BIRTH DATE, with the year, month, and day each being compared separately, and with the month and day being cross-compared where they had disagreed strongly on straight comparison,

GIVEN NAMES, i.e. the remainders of these, being straight-compared or cross-compared wherever the initials had already agreed or cross-agreed, and

RESIDENCE, i.e. the codes on the tax records for the province and locality of residence being compared with the place of the last medical examination as given on the WCB records.

In addition to the above identifiers, use was made of:

SEX, which was not weighted but served in this operation to identify apparent conflicts for visual inspection (although normally it would be required to agree so as to reduce the size of the file to be searched where the records initiating the searches all relate to males),

YEAR LAST KNOWN ALIVE, which served to exclude miners for whom there had been recorded contact since 1950,

PROBABLE DEATH STATUS, which served to identify apparent conflicts of the coded vital status on the pairs of records undergoing comparison, and

YEAR OF DEATH, which was used to single out instances in which there were conflicts, or agreements, relating specifically to the occurrence of deaths during the two tax years.

The comparison procedures are described in greater detail in Appendix B and its Table B.1.

In designing the comparison procedures, use was made of the frequencies of agreements, disagreements, and partial agreements, among the truly linkable pairs of records identified by the earlier searches based on the SIN number. These data provided the basis for calculating the odds in favour of a correct linkage, when employing the probabilistic approach. The procedures are described in Appendix B and the relevant data are illustrated in its Tables B.2 to B.6.

G. RESULTS FROM THE PROBABILISTIC LINKAGES

The test showed that in the absence of the SIN number the present work records could successfully search for and find the correctly matching tax records, using such identifiers as the names, birth dates, and the places of residence. It also showed that the level of accuracy was dependent on the extent to which the basic identifiers are represented on the work records.

The calculated degree of assurance of a correct linkage was high for the great majority of the matched pairs. Only 10 out of the 26,964 truly linkable pairs had theoretical odds of less than 1000-to-one in favour of a correct match (Table 8). Moreover, none of the 30,000 work records that initiated the searches were involved in false matches.

Much depended, however, on the completeness of the identifying information on the WCB records. Had the full date of birth been missing, or even just the month and day missing, many of the matched pairs would have had to be considered as "borderline links". Moreover, even the use of initials instead of full given names, or the absence of a middle initial, considerably reduced the degree of assurance of a correct link (see Table 9 and Appendix C).

In short, probabilistic linkages with the tax files are eminently feasible where the personal identifiers on the search records, other than the SIN number, are as complete and accurate as in the Ontario Workers' Compensation Board records. The level of success is entirely dependent, however, on the quality of the records initiating the searches. In particular, the full date of birth may be regarded as essential for such epidemiological "alive" follow-up.

H. DISCUSSION OF THE VALUE OF "ALIVE" FOLLOW-UP

The "alive" follow-up that has been demonstrated to be feasible will be of use mainly to those epidemiologists who are concerned with possible bias in their studies of the risks of death (see Appendices D and E). There are two main sources of bias. One of these is associated with errors and uncertainties in the death linkages. The other arises when there is an unidentified admixture of persons who are "lost to follow-up" among those presumed to be alive. This admixture of "seeming immortals" distorts most strongly the data on long term survival, especially when a cohort is followed to extinction. Linkages with a file of known living persons can substantially reduce both sources of bias.

Among the Ontario miners, the "lost to follow-up" group has been shown to constitute 3.4 per cent of the cohort. This is at a time when between 7 and 8 per cent of the cohort have died. The numbers "lost to follow-up" may be expected to increase somewhat as a cohort ages. However, the size of this group in the present cohort is not unlike that inferred for other similar cohort studies, due to migrations out of the country. Moreover, the figure may in the present instance be viewed more as an actual measurement, rather than as an estimate or inference.

Different investigators may view differently the merit of having such a measurement. What has been demonstrated here is that it can be obtained where the need exists. Moreover, the quality of such "alive" follow-up is seen to increase with the numbers of years of "alive" records that are searched. Thus, with linkage to only a single year of tax records, the "lost to follow-up" group would have been 4.9 per cent of the total cohort instead of the 3.4 per cent obtained with linkage to two years of tax records. Accuracy is likewise enhanced where a numeric "alive" search is possible for even a part of a cohort. This is because the data from the numeric linkages serve to refine the procedures for the probabilistic approach, based on names and such, that would be used where the SIN number is not available on the work record.

Not all cohort studies will require the increased refinement that can be achieved using the "alive" follow-up; but, for those that do, such refinement is eminently feasible. The income tax search provides the researcher with greater assurance in the quality of the death results being reported i.e. that deaths had not been missed in the computerized death search.

Table 1: Number of Miners' WCB Records Used for the "Alive" Follow-Up

(Ontario cohort of uranium and non-uranium miners for whom death searches had been carried out to the end of 1977.)

Reasons for exclusions from this file	Records exclu- ded	Records remain- ing
Total records in original file	-	50,279
Lacking SIN	20,070	30,209
Duplicate entries	76	30,133
Wrong SIN	133	30,000*

* These 30,000 miners' records with acceptable SIN numbers were used for matching against the 1977 and 1978 taxation files to verify the status of the individuals (i.e. as definitely alive in those taxation years, died in those years, or unaccounted for in the taxation files).

Table 2: Vital Status of the 30,000 Miners as Indicated by the Death Searches (including the manual resolutions of doubtful WCB-death links)

Status	Number of miners
Presumed alive (no acceptable death linkage found)	27,689
Uncertain (unresolved doubtful death linkage)	72
Probably dead (death linkage likely to be good)	137*
Definitely dead (death linkage positive)	2,102**

* In all subsequent tables the "probably dead" and "definitely dead" categories will be pooled and referred to simply as "dead".

** At the outset of the present study, an additional 44 miners had been recognized from the WCB death searches as having definitely died in 1978 and 1979. The numbers of "probably dead" and "uncertain" were unchanged at the end of 1979.

Table 3: Characteristics of the Taxation Files Used to Verify the Vital Status of the 30,000 Miners

NUMBERS OF RECORDS

1977 Taxation File	12,600,931 records
1978 Taxation File	14,320,408 records

PERSONAL IDENTIFIERS ON THE TAX RECORDS

Social Insurance Number
Names
Birth date

Sex
Marital status
Reporting province
Residence (province and locality)

Class of return*

* I.e. "normal", "dead", "immigrant/emigrant", "other".

Table 4: Results from Matching with the 1977 Taxation File Alone

(30,000 WCB records were matched, for exact agreement of SIN, against 12,600,931 taxation records.)

WCB status at end of 1977	<u>Taxation status on 1977 return</u>			Total
	alive	dead	no match	

Alive	26,196	55*	1,438**	27,689
Uncertain	44	1	27**	72
Dead	25*	215	1,999	2,239

Total	26,265	271	3,464	30,000
-------	--------	-----	-------	--------

* Apparent conflicts.

** Presumably "lost to follow-up".

Table 5: Results from Matching with Both the 1977 and 1978 Taxation Files, after Manually Resolving the Conflicts Between Tax Years

(30,000 WCB records were matched for exact agreement of SIN, against 12,600,931 tax records for 1977 and 14,320,408 for 1978; 38 conflicts between matching 1977 and 1978 taxation records were resolved manually.)

Status at end of 1977 (after 1977 tax link)	Taxation status 1977 (resolved with 1978)			
	alive	dead	no match	total
Alive	26,732	0	4*	26,736
Dead	9**	218	2,027	2,254
Lost to follow-up	0	5***	1,005	1,010
Total	26,741	223	3,036	30,000

* These 4 individuals were males, but their tax records were coded "female". As a result they were not picked up on the initial search of the tax records.

** These 9 individuals died prior to the end of 1977, but their tax records were wrongly coded "alive".

*** These 5 individuals were alive at the end of 1977, but their tax records were wrongly coded "dead".

NOTE 1: In 38 cases, believed by the WCB to be alive and later confirmed alive at the end of 1977, the 1977 and 1978 tax records both reported them to be dead. However, the deaths were found to have occurred early in 1978, i.e. prior to submission of the 1977 tax returns.

NOTE 2: An additional 281 deaths within the WCB cohort were found from the tax records to have occurred in 1978. Of that number, 63 have been confirmed by reference to the death records.

NOTE 3: One of the SIN numbers that is not among the 30,000 recognized in this study, was associated with a 1977 tax record coded "dead", which was followed by a 1978 tax record coded "alive". However, manual checking showed that SIN number to have been used by two different people, so it was excluded from the data reported in these tables.

Table 6: Final Disposition of the WCB-Tax Conflicts and Non-confirmations Shown in the Previous Table

WCB status 1977	Tax status	No. of rec.	Dispo- sition	Reason
a) live	dead	8	live	died early 1978 (not 1977)
b) live	dead	3	dead	death record found on microfiche
c) live	dead	5	lost to follow-up*	
d) live	no link	1,005	lost to follow-up	
e) uncert. live		43	live	no reason to doubt tax status
f) uncert. dead		1	dead	no reason to doubt tax status
g) uncert. no link		27	dead**	
h) dead	live	16	live	live on both 1977 and 1978 tax file
i) dead	live	8	dead	estates did not report deaths
j) uncert. live		1	dead	estate did not report death

* These are probably dead but were assigned to "lost to follow-up" because the death records could not be found on microfilm, and the deaths may well have occurred outside the country.

** These are judged "dead" rather than "lost to follow-up" because of two circumstances:

a) a possible, if dubious, death link had been found, and

b) most living cohort members did file tax returns, in a ratio of roughly 26 : 1.

Table 7: Vital Status as Judged Prior to and Following the Taxation Searches and the Manual Resolutions -- Summary

(See Appendix A and its Tables A.1 and A.2 for detail.)

Previous and final assessments of the WCB records	Inferred vital status				Total
	alive	uncer- tain	dead	lost to follow- up	

After death link	27,689	71	2,240	-	30,000
After 77 tax file	26,240	80	2,215	1,465	30,000
After 77/78 tax link	26,711	42	2,215	1,032	30,000
After final resolution	26,736	-	2,254	1,010	30,000

Table 8: Frequencies of Good Matches, False Matches and Non-Matches from the Probabilistic "Alive" Searches

(Only the "best" match for each WCB record is considered here.)

<u>Total weight</u>			<u>Linkable WCBs</u>		<u>Unlinkable WCBs</u>		<u>Total WCB records</u>
odds ratio*	wt	range	good matches	bad matches	good matches	bad matches	
211+	105	and over	26,955	-	-	-	26,955
210	95	to 104	2	-	-	-	2
29	85	94	3	-	-	-	3
28	75	84	2	-	-	-	2
27	65	74	-	-	-	-	-
26	55	64	-	-	-	-	-
25	45	to 54	1	-	-	-	1
24	35	44	1	-	-	-	1
23	25	34	-	-	-	-	-
22	15	24	-	-	-	-	-
21	5	14	-	-	-	-	-
20+	0	to 4	-	-	-	-	-
20-	- 5	- 1	-	-	-	-	-
2-1	-15	to - 6	-	-	-	-	-
2-2	-25	-16	-	-	-	-	-
2-3	-35	-26	-	-	-	-	-
2-4	-45	-36	-	-	-	-	-
2-5	-55	-46	-	-	-	-	-
2-6	-65	to -56	-	-	-	-	-
2-7	-75	-66	-	-	-	-	-
2-8	-85	-76	-	-	-	-	-
2-9	-95	-86	-	-	-	-	-
2-10	-105	-96	-	-	-	-	-
2-11	-115	and below	-	-	-	-	-
no match			(-)	(3,036)	(-)	-	3,036
Total			26,964**	- (3,036)	-	-	30,000

* Odds in favour of a correct link: 20 equals 50:50 odds point.

** Includes 57 links in which the surname NYSIIS code disagree.

Table 9: Accuracy of the Probabilistic "Alive" Searches with Different Amounts of Identifying Information

Identifiers present on WCB record		Number of record pairs	Percentage distribution by theoretical odds in favour of a correct linkage			
1st name	2nd name		21 to 210	211 to 220	221 to 230	231 and up
FULL DATE OF BIRTH PRESENT ON BOTH RECORDS						
full	full	1,937	.05%	.21%	1.60%	98.14%
full	init	9,770	.01	.06	1.93	98.00
full	--	14,760	.05	.31	5.22	94.42
init	init	175	-	1.14	12.57	86.29
-all other-		322	-	.62	11.80	87.58
Combined		26,964	.03%	.22%	3.89%	95.85%

MONTH AND DAY OF BIRTH MISSING

The odds in favour of linkage would be decreased by a factor of more than 28, i.e. by almost 1/365.

YEAR, MONTH AND DAY OF BIRTH MISSING

The odds in favour of linkage would be decreased by a factor of 214, i.e. by about 1/1600.

NOTE 1. The percentages are based on the numbers of good links as determined earlier from the numerical "alive" searches using the SIN number. For details of the counts see Appendix Table C.1.

NOTE 2. The linked pairs include 280 (1.04 per cent) in which the surname NYSIIS codes agreed but the detailed spelling did not, and 57 pairs (0.21 per cent) in which the NYSIIS codes disagreed. The latter would have been missed in a probabilistic match based on a search of the tax file in surname NYSIIS sequence.

ACKNOWLEDGEMENTS

The authors would like to thank the Ontario Workers' Compensation Board and the Ontario Ministry of Labour for use of their data files and for their contribution to this study.

APPENDIX A

VITAL STATUS AS JUDGED PRIOR TO AND FOLLOWING THE TAXATION SEARCHES AND MANUAL RESOLUTIONS -- DETAIL

The taxation records assist in various ways in the resolution of uncertainties concerning vital status:

- by the fact that there is, or is not, a tax return for a given individual in a given year,
- by identifying the tax returns that relate to estates,
- by identifying the tax returns that relate to immigrants and emigrants.

These aids have been used in the manual resolutions of the vital status of the Ontario miners. The final resolutions have already been summarized in Table 7 of the main text. Further details are given in Appendix Table A.1.

The associations between the taxation status on the two successive years is revealing. A breakdown which recognizes immigrant and emigrant status in the earlier taxation year, and the proportions filing returns in the following year, does show the expected correlation between the two. The data are as follows:

1977 tax status	Proportion submitting 1978 tax returns
1) Normal (alive)	25,712 / 26,154 = 98.3%
2) Dead	39 / 271 = 14.4
3) No tax return	433 / 3,464 = 12.5
4) Immigrant/emigrant	3 / 3 = 100.0%
5) Immigrant	9 / 10 = 90.0
6) Emigrant	12 / 23 = 52.2
7) Other	67 / 76 = 88.2
Combined	26,275 / 30,001 = 87.6%

The conclusions that may be drawn are fairly obvious:

- The majority of persons submitting tax returns in given year will do so again in the following year (98%).
- The estates of deceased persons sometimes submit tax returns in two successive years (14%).
- Immigrants usually continue to submit tax returns, even when they later emigrate (92%).
- About half of the taxpayers who emigrate submit returns the year after emigration (52%).

Clearly, two years of tax returns are more revealing of the vital status than just a single year. It follows that successive years of "alive" follow-up, using the tax files over longer periods, would provide even more precise indications of the survival history of an epidemiological cohort.

Appendix Table A-1: Vital Status as Judged Prior to and Following the Taxation Searches and Manual Resolutions -- Detail

Previous Total assessment of the WCB records	Revised assessment				
	alive	uncer- tain	dead	lost to follow- up	
AFTER 1977 TAX LINKAGE ONLY					
Alive	26,196	55	-	1,438	27,689
Uncertain	44	-	1	27	72
Dead	-	25	2,214	-	2,239
Total	26,240	80	2,215	1,465	30,000
AFTER 1977 AND 1978 TAX LINKAGES (with tax-tax conflicts resolved)					
Alive	26,667	17	-	1,005	27,689
Uncertain	44	-	1	27	72
Dead	-	25	2,214	-	2,239
Total	26,711	42	2,215	1,032	30,000
AFTER RESOLVING WCB-TAX CONFLICTS					
Alive	26,676	-	3	1,010	27,689
Uncertain	44	-	28	-	72
Dead	16	-	2,223	-	2,239
Total	26,736	-	2,254	1,010	30,000
COMBINED					
After death link	27,689	71	2,240	-	30,000
After 77 tax link	26,240	80	2,215	1,465	30,000
After 78 tax link	26,711	42	2,215	1,032	30,000
After resolution	26,736	-	2,254	1,010	30,000

PROCEDURES FOR THE PROBABILISTIC LINKAGES -- DETAIL

Wherever a miner's record was linked to a taxation record using the SIN number, a probabilistic linkage was carried out as well. This served two purposes. First, it singled out for visual inspection any of the numeric linkages that seemed questionable. Second, it provided data that were needed to improve the probabilistic linkage procedure.

The identifiers that were compared when doing the probabilistic linkages are indicated in Appendix Table B.1. The manner of their comparison, i.e. the levels of agreement, disagreement, and partial agreement that were recognized when calculating the odds in favour of or against a correct linkage, are indicated in Appendix Table B.2. The "weights" shown in this latter Table are simply the logarithms of the "odds" associated with a particular outcome from a given comparison. To be more precise, they are the logarithms to the base 2, multiplied by 10 to avoid decimals.

There is a positive component and a negative component to the weight for each such outcome, whether it is an agreement, a disagreement, or a partial agreement. Taken together, these constitute a net weight. The logic is quite simple. The positive component is based on the frequency of the particular outcome among unlinkable pairs of records brought together by chance; outcomes that are rare among unlinkable pairs will tend to argue strongly in favour of linkage when they do occur. The negative component is based on the frequency the same outcome among truly linkable pairs of records; outcomes that are rare among these will, of course, tend to argue strongly against a correct linkage. The net weight reflects the balance between the two; it is simply the logarithm of the ratio between the two frequencies (i.e. the log to the base 2, multiplied by 10). The formulae are given below.

For the comparison of any identifier, and for any OUTCOME from that comparison, defined in any way (e.g. agreement, disagreement, partial agreement, etc.) the weights will be as follows:

$$\text{pos component} = 10 \times \log_2 [1/(\text{freq in unlinkable pairs})] \quad [1]$$

$$\text{neg component} = 10 \times \log_2 (\text{freq in linked pairs}) \quad [2]$$

$$\text{net weight} = \frac{10 \times \log_2 (\text{freq in linked pairs})}{10 \times \log_2 (\text{freq in unlinkable pairs})} \quad [3]$$

$$\text{net weight} = \text{pos component} + \text{neg component} \quad [4]$$

The application of these formulae to the calculation of weights for comparisons of geographic information are of special interest. Specifically, the place of the last medical examination from the work records was compared with the place of residence as given on the death records. The data on which the weights were calculated for comparisons of the province, and of the locality when Ontario was the province on both records, are given in Appendix Tables B.3 and B.4.

Province of residence, and locality of residence within the province, are coded on the 1977 Taxation records and may be compared with the place of the last medical examination as given on the Ontario Workers' Compensation Board records. The codes for the localities differ somewhat but the comparisons are still possible and show strong correlations that can be weighted.

Initially, in order to carry out probabilistic linkages on files that have not been previously linked, weights from other linkage operations are frequently used. Where these are found to be inappropriate for the current linkage operation, more suitable weights may be derived from the new set of linked pairs coming out of that operation, and from a new set of unlinkable pairs that have been randomly matched from the current files. A weight update may result; and, if this alters the linkage status of any substantial proportion of the pairs in the linked file, the negative components can be re-calculated. Conceptually, this kind of fine tuning of the weights can be an iterative operation, which is terminated when no further worthwhile improvement is likely.

The weights for the above comparisons of geographic information have been based on the odds ratio formula and its components. The approach has worked well, both for the geographic information and also for the comparisons of the year, month and day of birth.

The same approach would be correctly applicable also to the calculation of weights for comparisons of name information, but has not actually been used here. Instead, these weights have been borrowed from previous linkage operations. They differ greatly from the values that would have been obtained using the odds ratio formula, and appear to be in need of revision. Revised weights have been calculated that might profitably be substituted in the future are given in Appendix Table B.5 to B.7. Such comparisons are important as indicating inconsistencies in weighting procedure, and the steps by which they may be avoided.

The weights for the name comparisons illustrated in the Tables are all non-specific for the "value" of the name. At a later stage in the linkage procedure, the weights for agreements and partial agreements of names are changed to become "value specific". Look-up tables are required, which indicate how common or how rare particular names are. The rare names, when they agree, will carry more than average positive weight in favour of a correct linkage. Conversely, agreements of common names will carry less than average positive weight. For reasons of economy in the use of computer time, this refinement is generally applied after most of the unpromising candidate pairs for linkage have been eliminated.

Appendix Table B-1: Identifiers Compared Between the Workers' Compensation Board Records and those of the Taxation Files

Identifiers compared	Comparison name	WCB record field	Tax record field	Footnotes
1 Surname and its phonetic code	SURNAME	NYSIIS SURNAME	NYSIIS SURNAME	1
2 Surname weight class	NCLASS	WTCLASS	SURNAME	1
3 First initial	INIT11	GIVEN1	GIVEN1	
4 Second initial	INIT22	GIVEN2	GIVEN2	
5 Initial cross-compared (1st-2nd)	INIT12	GIVEN1	GIVEN2	2
6 Initial cross-compared (2nd-1st)	INIT21	GIVEN2	GIVEN1	2
7 Birth year (degree of discrepancy)	YRDIFF	BIRTHYR	BIRTHYR	
8 Birth month (degree of discrepancy)	MNDIFF	BIRTHMN	BIRTHMN	
9 Birth day (degree of discrepancy)	DYDIFF	BIRTHDY	BIRTHDY	
10 Birth month-day cross-comparison	XBDATE	BIRTHMN BIRTHDY	BIRTHMN BIRTHDY	3
11 First given name (degree of discrep)	GIVEN11	GIVEN1	GIVEN1	
12 Second given name (degree of discrep)	GIVEN22	GIVEN2	GIVEN2	
13 Given cross-comparison (1st-2nd) (degree of discrepancy)	GIVEN12	GIVEN1	GIVEN2	4
14 Given cross-comparison (2nd-1st) (degree of discrepancy)	GIVEN21	GIVEN2	GIVEN1	4
15 Sex	SEX	SEX	SEX	5
16 Province of residence	PRRES	PLLEX	PRRES	6
17 Locality of residence	LOCALTY	PLLEX	CNTY	6
18 Year last known alive	YRLKA	YRLKA	SRCODE	7
19 Vital status	DTPROB	DTPROB	SRCODE	7
20 Death year	DTHYR	DTHDATE	SRCODE	7

Footnotes to Appendix Table B-1

1. For a given level of agreement, any one of five different weights may be assigned initially, depending on how rare or common the surname is (or the portion of it that agrees).
2. Initials are only cross-compared where there is no straight agreement.
3. Birth month and day are only cross-compared where there is no straight agreement.
4. The remainders of the given names are only cross-compared where the initials have already been found to cross-agree.
5. Since all of the Ontario miners are males, no weight is attached to agreement or disagreement of SEX; but disagreement will result in a visual check to determine if there is a miscoding.
6. The locality of residence is only compared where the tax record shows the province of residence to be Ontario.
7. These comparisons are not weighted, but are used in the visual checking.

Appendix Table B-2: Weights Assigned to the Comparison Outcomes at the Time of the First Production (at the Frequency Weighting Phase).*

(All weights are non-specific for the particular identifier value.)

Outcome level		Pos com- pon- ent	Neg com- pon- ent	Net wt*
1	SURNAME			
	missing			0
	full agreement (up to 7 characters)	195	-1	194
	partial 2 (4 to 6 characters agree)	195	-53	142
	partial 3 (only NYSIIS code agrees)	195	-71	124
	disagree (NYSIIS code disagrees)			-66
2	NCLASS			
	(Look-up tables are used to obtain weights for 3 levels of agreement, and 5 levels of specificity of the particular name or portion of the name that agrees.)			
3	INIT11			
	missing			0
	agreement	38	0	38
	disagreement			-46
4	INIT22			
	missing			0
	agreement	38	-1	37
	disagreement			-38
5	INIT12			
	(where no straight agreement of either initial)			
	missing			0
	agreement	38	-1	37
	disagreement			0
6	INIT21			
	(where no straight agreement of either initial)			
	missing			0
	agreement	38	-3	35
	disagreement			0
7	YRDIFF			
	missing			-10
	full agreement	71	-2	69
	partial 2 (difference = 1)	60	-34	26
	partial 3 (difference = 2 to 3)	51	-49	2
	partial 4 (difference = 4 to 5)	41	-71	-30
	partial 5 (difference = 6 to 9)	31	-82	-51
	disagreement (difference = 10+)			-92
8	MNDIFF			
	missing			-10
	full agreement	36	-2	34
	partial 2 (difference = 1)	27	-64	-37
	partial 3 (difference = 2)	29	-76	-47
	partial 4 (difference = 3)	30	-88	-58
	disagreement (difference = 4+)			-69

Continued on next page

Appendix Table B-2 (continued)

9	DYDIFF	missing			-10
		full agreement	49	-4	45
		partial 2 (difference = 1)	40	-54	-14
		partial 3 (difference = 2 to 3)	30	-65	-35
		partial 4 (difference = 4 to 9)	15	-60	-45
		disagreement (difference = 10+)			-56
10	XBDATE	(no straight agreement on either month or day)			
		both cross-agree	0	+90	90
		XDATE 2 (WCB month = TAX day)	0	+53	53
		XDATE 2 (WCB day = TAX month)	0	+37	37
		no cross agreement			0
11	GIVEN11	(first initials already agree)			
		rest of given name missing			0
		full agreement (up to 7 characters)	119	-1	118
		partial 2 (first 4 agree + missing)	119	-1	118
		partial 3 (first 4 agree + disagree)	119	-55	64
		partial 4 (first 3 agree + missing)	119	-1	118
		partial 5 (first 3 agree + disagree)	119	-57	62
		partial 6 (first 2 agree + missing)	119	-1	118
		partial 7 (first 2 agree + disagree)	119	-58	61
		partial 8 (first 1 agrees + disagree)	0	-59	-59
		other			0
12	GIVEN22	(second initials already agree)			
		rest of given name missing			0
		full agreement (up to 7 characters)	119	-1	118
		partial 2 (first 4 agree + missing)	119	-1	118
		partial 3 (first 4 agree + disagree)	119	-55	64
		partial 4 (first 3 agree + missing)	119	-1	118
		partial 5 (first 3 agree + disagree)	119	-57	62
		partial 6 (first 2 agree + missing)	119	-1	118
		partial 7 (first 2 agree + disagree)	119	-58	61
		partial 8 (first 1 agrees + disagree)	0	-53	-53
		other			0
13	GIVEN12	(1st-2nd initials already cross-agree)			
		rest of given name missing			0
		full agreement (up to 7 characters)	119	-8	111
		partial 2 (first 4 agree + missing)	119	-8	111
		partial 3 (first 4 agree + disagree)	119	-47	72
		partial 4 (first 3 agree + missing)	119	-8	111
		partial 5 (first 3 agree + disagree)	119	-50	69
		partial 6 (first 2 agree + missing)	119	-8	111
		partial 7 (first 2 agree + disagree)	119	-26	93
		partial 8 (first 1 agrees + disagree)	0	-24	-24
		other			0

Continued on next page

Appendix Table B-2 (continued)

14	GIVEN21 (2nd-1st initials already cross-agree)			
	rest of given name missing			0
	full agreement (up to 7 characters)	119	-8	111
	partial 2 (first 4 agree + missing)	119	-8	111
	partial 3 (first 4 agree + disagree)	119	-47	72
	partial 4 (first 3 agree + missing)	119	-8	111
	partial 5 (first 3 agree + disagree)	119	-50	69
	partial 6 (first 2 agree + missing)	119	-8	111
	partial 7 (first 2 agree + disagree)	119	-26	93
	partial 8 (first 1 agrees + disagree)	0	-24	-24
	other			0
15	SEX (no weights assigned)			
16	PRRES (WCB always equals Ontario)			
	missing			0
	full agreement (Ontario on death record)	14	-2	12
	partial 2 (PEI on death record)	77	-116	-39
	partial 3 (NS on death record)	48	-78	-30
	partial 4 (NB on death record)	52	-75	-23
	partial 5 (PQ on death record)	20	-45	-25
	partial 6 (Man on death record)	44	-60	-16
	partial 7 (Sask on death record)	47	-73	-26
	partial 8 (Alta on death record)	35	-63	-28
	partial 9 (BC on death record)	31	-53	-22
	partial 10 (Yuk on death record)	91	-84	7
	partial 11 (NWT on death record)	103	-100	3
	partial 12 (Nfld on death record)	88	-104	-16
	other			0
17	LOCALITY (where province is Ontario on both records)			
	(* = nearest equivalent on TAX record.)			
	missing			0
	full agreement (Timmins on WCB, plus *)	74	-26	48
	partial 2 (Kirkland Lake on WCB, plus *)	97	-40	57
	partial 3 (Sudbury on WCB, plus *)	54	-12	42
	partial 4 (Fort William on WCB, plus *)	51	-34	17
	partial 5 (Elliot Lake on WCB, plus *)	100	-35	65
	partial 6 (Bancroft on WCB, plus *)	30	0	30
	partial 7 (North Bay on WCB, plus *)	77	-53	24
	disagree (= all other outcomes)			0
18	YRLKA (no weights assigned)			
19	DTPROB (no weights assigned)			
20	DTHYR (no weights assigned)			

Notes Regarding Table B-2

The weights in this table relate to those estimates available at the time of the first production. Further refinements to a number of weight sets may be needed for future studies. These are shown in Tables B3-B6.

Notes regarding specific outcomes are as follows.

1. SURNAME. The weights for full agreement, partial 2, partial 3 are the maximum weights available. Look-up tables were used to obtain specific weights for particular names or portions of the name that agree. These weights may need revision in future studies (see Table B-6).
2. NCLASS. This was used only at the basic compare phase, and is set to zero at the frequency weighting phase.
- 3-6. INIT 11, INIT 22, INIT 12, INIT 21. These were the weights used in the basic compare phase. In the frequency weighting phase, these weights are used if the values are not found in the weight set.
- 7-9. YRDIFF, MNDIFF, DYDIFF. The positive component weight estimate came from an earlier linkage and were modified. For future linkages of these files, use the improved estimates available in Table B-5, which were generated using a file of NONLINKS.
- 11-14. GIVEN 11, GIVEN 22, GIVEN 12, GIVEN 21. The weights for full agreement and partial agreement are the maximum weights available. Look-up tables are used to obtain specific weights for particular names or portions of the names that agree. For future studies, see the weight in Table B-7.
16. PRRES. The positive component of the weight was derived from a file of NONLINKS (see also Table B-3 for details). The negative component estimate came from the linked pairs.
17. LOCALITY. The positive component of the weight was derived from a file of NONLINKS (see also Table B-4 for details). The positive component estimate came from the linked pairs.

Appendix Table B-3: Weights for Comparisons with PROVINCE OF RESIDENCE as Given on the 1977 Taxation File -- calculated after the linkage versus actually used

(Province of last medical examination on the Ontario WCB records is always "Ont".)

Outcome (Tax resid- ence)	Counts		Percentages		Weights (calc)			Weights (used)		
	non- links	good links	non- links	good links	pos wt	neg wt	net wt	pos wt	neg wt	net wt
1 Ont	11275	45021	38.5	88.2	14	-2	12	14	-2	12
2 PEI	139	17	.4	.0	77	-116	-39	77	-116	-39
3 NS	1021	231	3.5	.5	48	-78	-30	48	-78	-30
4 NB	788	286	2.7	.6	52	-75	-23	52	-75	-23
5 Que	7489	2176	25.6	4.3	20	-45	-25	20	-45	-25
6 Man	1403	776	4.8	1.5	44	-60	-16	44	-60	-16
7 Sask	1137	333	3.9	.7	47	-73	-26	47	-73	-26
8 Alta	2546	644	8.7	1.3	35	-63	-28	35	-63	-28
9 BC	3311	1292	11.3	2.5	31	-53	-22	31	-53	-22
10 Yuk	55	156	.2	.3	91	-84	7	91	-84	7
11 NWT	23	49	.1	.1	103	-100	3	103	-100	3
12 Nfld	66	36	.2	.1	88	-104	-16	88	-104	-16
Combined	29253	51017	100.0	100.0						

Appendix Table B-4: Weights for Comparisons with LOCALITY OF RESIDENCE as given on the 1977 Taxation File -- calculated after the linkage versus actually used

(Province of last examination as given on the Ontario WCB records is always "Ont". The locality of the last medical examination varies within the province.)

Outcome (see detail below)	Counts		Percentages		Weights (calc)			Weights (used)		
	non- links	good links	non- links	good links	pos wt	neg wt	net wt	pos wt	neg wt	net wt
1 Timmins	68	7298	.6	16.2	74	-26	48	74	-26	48
2 Sudbury	14	2865	.1	6.4	97	-40	57	97	-40	57
3 Sudbury	271	19646	2.4	43.6	54	-12	42	54	-12	42
4 Thunder	323	4201	2.9	9.3	51	-34	17	51	-34	17
5 Elliot	11	4002	.1	8.9	100	-35	65	100	-35	65
6 Bancroft	-	534	.0	1.2	145*	-64*	81*	30*	0*	30*
7 North B.	54	1158	.5	2.6	77	-53	24	77	-53	24
8 Other	10534	5317	93.4	11.8	1*	-31*	-30*	*	*	0*
Total	11275	45021	100.0	100.0						

* NOTE: In these two cases there is a substantial discrepancy between the weights calculated after the linkage, and those actually used for linkage. The calculated weights are believed to be the correct values. Where a count is zero, the conservative approach is to assume a count of 0.5, as in Outcome number 6.

(The details of the localities compared are given on the next page.)

Appendix Table B-4 (continued)

Details of the Localities Compared

Outcome number	Details of the localities compared
1	WCB: Timmins (South Porcupine, Tisdale) Tax: Timmins, Cochrane, Porcupine, Schumacher
2	WCB: Kirkland Lake Tax: Kirkland Lake, Timiskaming
3	WCB: Sudbury (Coppercliff, Falconbridge, Nickel) Tax: Sudbury County and Municipality, Coppercliff, Nipissing
4	WCB: Thunder Bay (Port Arthur, Fort William) Tax: Thunder Bay County and Municipality, Algoma, Rainy River, Kenora, Patricia
5	WCB: Elliot Lake (Blind River) Tax: Algoma, Sudbury County and Municipality, Coppercliff
6	WCB: Bancroft Tax: Hastings County
7	WCB: North Bay Tax: Algoma, Nipissing, Timiskaming
8	All other combinations

Appendix Table B-5: Weighting for Comparison of birth years

Outcome level	Counts		Percentages		Weights (calc)			Weights (used)		
	non-links	good links	non-links	good links	pos wt	neg wt	net wt	pos wt	neg wt	net wt
1 Full agreement	435	43870	1.5	85.8	61	-2	59	71	-2	69
2 + or - 1 year	951	4544	3.2	8.9	50	-35	15	60	-34	26
3 + or - 2 to 3 years	1861	2004	6.2	3.9	40	-47	-7	51	-49	2
4 + or - 4 to 5 years	1946	327	6.5	0.6	39	-73	-34	41	-71	-30
5 + or - 6 to 9 years	3791	179	12.7	0.3	30	-82	-52	31	-82	-51
6 + or 1-10 & more	20915	267	69.9	0.5	5	-76	-71			-92
Total	29899	51191	100.0	100.0						

Appendix Table B-6: Weights in Need of Further Study -- SURNAME and SURNAME NYSIIS

Outcome level	Counts		Percentages		Weights (calc)			Weights (used)		
	non-links	good links	non-links	good links	pos wt	neg wt	net wt	pos wt	neg wt	net wt
1 (7 agr)	15	48781	.1	95.3	110	-1	109*	195	-1	194*
2 (4 agr)	1	732	0	1.4	149	-61	88*	195	-53	142*
3 (NYS agr)	10	565	0	1.1	115	-65	50*	195	-71	124*
(NYS dis)	29873	1111	99.9	2.2	0	-55	-55			-66
Total	29899	51189	100.0	100.0						

*NOTE: Asterisks indicate major discrepancies between the weights calculated after the linkage, and those actually used for the linkage.

Appendix Table B-7: Weights in Need of Further Study -- FIRST GIVEN NAMES

(Outcome levels are rearranged for clarity.)

Outcome level	Counts		Percentages		Weights (calc)			Weights (used)		
	non-links	good links	non-links	good links	pos wt	neg wt	net wt	pos wt	neg wt	net wt
1 (agree)	230	42400	10.8	87.9	32	-2	30*	119	-1	118*
2 (4 + M)	4	860	.2	1.8	90	-58	32*	119	-1	118*
4 (3 + M)	4	371	.2	.8	90	-70	20*	119	-1	118*
6 (2 + M)	1	28	.0	.1	110	-108	2*	119	-1	118*
3 (4 + D)	19	1650	.9	3.4	67	-49	18*	119	-55	64*
5 (3 + D)	42	956	2.1	2.0	56	-57	-1*	119	-57	62*
7 (2 + D)	327	914	16.0	1.9	26	-57	-31*	119	-58	61*
8 (1 + D)	1428	1057	69.8	2.2	5	-55	-50	0	-59	-59
Total	2045	48236	100.0	100.0						

*NOTE: Asterisks indicate major discrepancies between the weights calculated after the linkage, and those actually used for the linkage.

ACCURACY OF THE PROBABILISTIC LINKAGES -- DETAIL

The accuracy of the probabilistic linkages is dependent to a considerable degree on the availability of the GIVEN NAMES, as distinct from just the INITIALS, even when the full DATE OF BIRTH is present. This is shown most clearly by a breakdown of the levels of assurance of a correct linkage (for the actual counts see Appendix Table C.1).

As would be expected, the proportion of correctly matched pairs falling into the higher levels of assurance (i.e. the higher weight ranges) is greater where there is a middle name or initial than where there is not. Moreover, it is better to have a middle name than to have just a middle initial.

The present test made use only of those WCB records and taxation records that were known in advance to be potentially linkable with each other. As a result, the Table does not tell us how many truly unlinkable pairs of records would fall into the various levels of assurance, in a linkage operation based on all WCB and taxation records. The extent of the overlap, and the optimum position for a threshold in the scale of assurances, have not been measured. However, subjectively and from one's memory of other linkage operations, it is believed that the overlap would be small. In other words, with an optimum threshold, relatively few false positive links and false negatives would be expected, and most of these would involve pairs in which the full given names were lacking on at least one of the records.

In the absence of the full date of birth, however, such a search of the taxation files would be largely unreliable and of much less practical use.

Appendix Table C-1: Accuracy of the Probabilistic "Alive" Searches with Different Amounts of Identifying Information -- Detail

Weight	Record pairs by weight and by presence or absence range of given name or initials (see code below)					
	FF	FI	F-	II	OTHR	Total
705-804	8	-	-	-	-	8
605-704	69	32	3	-	-	104
505-604	418	981	1,422	-	8	2,829
405-504	1,074	5,803	6,277	57	120	13,331
305-404	332	2,758	6,235	94	154	9,573
205-304	31	189	770	22	38	1,050
105-204	4	6	46	2	2	60
5-104	1	1	7	-	-	9
Total	1,937	9,770	14,760	175	322	26,964

NOTE 1: FF = full first and second given names on both records
 FI = full first name plus middle initial on both records
 F- = full first name, no middle initial, on both records
 II = first and second initials, on both records
 OTHR = all other combinations

NOTE 2: A weight range of 5 to 104 is taken to represent a range of odds, in favour of a correct link, of 21 to 210. Similarly, a weight range of 105 to 204 represents an odds range of 211 to 220, and so on.

SOME ESSENTIAL FEATURES OF MORTALITY FOLLOW-UP STUDIES AMONG INDUSTRIAL POPULATIONS

This Appendix describes essential features of complete historic cohort studies among industrial workers and is based on description by Symons and Taulbee¹. These essential features are of interest in designing and evaluation long-term health studies of persons exposed to low levels of radiation.

The cornerstone for most historic cohort mortality studies is in determining the amount of risk of death for the workforce over an observation period. This evaluation requires the combined information on the number of cohort members, the period of observation, and the expected deaths for all causes or a specific cause are calculated. The ratio of the observed deaths to the expected deaths, possibly multiplied by 100, and reported as a percentage, is termed the standardized mortality ratio. The standardization refers to the specific death ratio used from a comparison population.

The cohort of workers is followed and the vital status determined as of the end of the observation period (see Figure 1). The percentage of death certificates obtained for those known deaths, the proportion "lost to follow-up" (that is the fraction of individuals identified at the outset but with unknown vital status at the end of the study) and the proportion of the cohort enumerated should be reported as part of the study results. These proportion bear directly on the reliability of the conclusions derived since they describe the amounts of missing data in the study.

The proportion of death certificates available for those known decedents is directly calculable. With use of the Canadian Mortality Data Base, all deaths occurring in Canada are available. Deaths of Canadian residents where death has occurred in the United States, are also available, although reporting is believed to be incomplete. Using an "alive" and "death" search for the entire cohort, one can evaluate the quality of the effort can be directed at resolving these doubtful linkages.

The proportion of the enumerated cohort that are "lost to follow-up" is directly calculable and is an upper bound on the number of unknown deaths.

The proportion of the defined cohort that is enumerated is not really directly estimable, but describing the diligence of cross checking of various sources, internal to the industry and external ones (for example, local union and provincial and/or federal government offices), to ensure completeness of the enumerates of all cohort numbers enhance the study validity. The exact method to be employed would be expected to vary with each situation.

References to Appendix D

1. Symons MJ, Taulbee JD. Statistical evaluation of the risk of cancer mortality among industrial populations. In: O'Connell RG ed. Statistical methods for cancer studies. Marcel Dekker, Inc. 270 Madison Avenue, New York 1984: 25-90.

DEALING WITH PERSONS "LOST TO FOLLOW-UP"

This Appendix describes ways of dealing with persons "lost to follow-up" in the analysis of industrial populations, and is based on a description given by Monson¹ regarding methods used in studies conducted in the United States.

It is desirable to determine the vital status at the end of the potential period of follow-up for all persons in a study cohort. Usually, however, this is not possible. There are always persons who are not traceable. Monson indicates that such persons must be considered to be under follow-up only for the known period of observation, say the years during which they worked. In the United States, if some mass form of follow-up is used, such as the Social Security Administration, it may not be known which persons not known to be dead are still alive and which are lost. In this event, all nondead persons usually are considered to be alive at the end of follow-up.

If an individual follow-up has been carried out, four basic assumptions can be made about persons who are "lost to follow-up".

1. They can be eliminated from the study.
2. They may be assumed to be deceased at the end of follow-up.
3. They may be assumed to be alive at the end of follow-up.
4. They may be assumed to be lost at the date of loss.

Assumption one is usually not advisable. All persons once they are entered into follow-up are at risk of dying and contribute to the number of deaths expected in the study cohort. If they are excluded, the expected number of deaths will be artificially low. However, if the observed deaths are fewer than the expected, even with the exclusion of the lost person, any error introduced leads to a conservative interpretation of the data. The cohort only appears to be less healthy than it actually is.

Assumption two is usually not realistic. If the observed deaths are fewer than those expected even with this extreme assumption, a conservative interpretation of the data again results.

Assumption three is the most conservative assumption if an excess of some cause of death is due to that cause probably have not been identified, and the number of deaths expected is maximal. If even with those two opposing errors in measurement an excess is seen, it is an underestimate of the true excess. However, if no excesses of specific causes are seen under assumption three, than a reanalysis should be made using assumption four.

Assumption four assumes that the mortality of persons who were lost does not differ from the mortality of persons who continue under follow-up.

Assumption four and assumption three provide reasonable lower limits and upper limits as to the range of the numbers of deaths expected from various causes.

Monson states that in presenting data as to the percentage of follow-up, the usual method is to give the percentage of persons who were followed until death or until the end of the study. This method, he states, tends to be conservative. Most persons who become "lost to follow-up" are followed for some number of years and contribute to the person-years and the expected number of deaths. Instead of computing the percentage of persons followed, one could compute the percentage of the maximum number of person-years or the percentage of the maximum number of expected deaths from all causes. Either of these percentages will usually be greater than the percentages of persons followed. For example, in a study of rubber workers only 69% of white male union members were followed until death or end of follow-up through company record². However, the follow-up through these records represented 85% of the potential maximum person-years of follow-up, and 82% of the maximum number of expected deaths from all causes. In describing follow-up data, Monson suggests all three of these percentages could be presented.

References to Appendix E

1. Monson RR. Occupational epidemiology CRC Press Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida, 33431 1980.
2. Monson RR and Nakano KK. Mortality among rubber workers. I. White male union employees in Akron, Ohio. Am. J. Epidemiology, 103, 284, 1976.

STATISTICS CANADA POLICY ON RECORD LINKAGE

Statistics Canada has developed a policy on record linkage. This project satisfied the prescribed review and approval process.

Record linkage is an important technique in the development, production, analysis, and evaluation of statistical data. The record linkage policy is about the methods and procedures to be used to identify, on a case by case basis, the balance between the public good to be served by each record linkage proposed and the potential invasion of privacy.

The policy statement is as follows.

Statistics Canada will undertake record linkage activities only if all the following conditions are satisfied:

- . the purpose of the record linkage activity is statistical/research and is consistent with the mandate of Statistics Canada as described in the Statistics Act; and
- . the products of the record linkage activity will be released only in accordance with the confidentiality provisions of the Statistics Act and with any applicable requirements of the Privacy Act; and
- . the record linkage activity has demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option; and
- . the record linkage activity will not be used for purpose that can be detrimental to the individuals involved and the benefits to be derived from such a linkage are clearly in the public interest; and
- . the record linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs; and
- . the linkage satisfies a prescribed review and approval process.